Multidimensional data analysis

Joe Donovan

MPI of Neurobiology

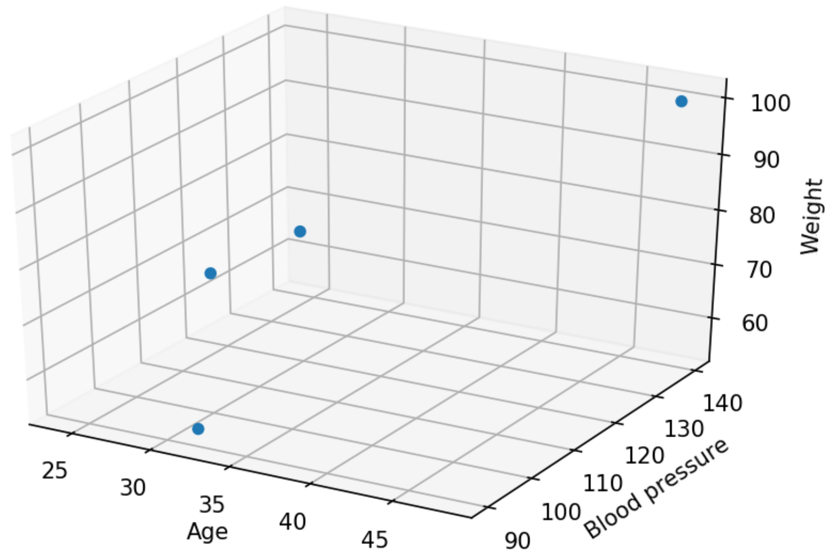
max-planck-institut für
neurobiologie

Most data is multidimensional

- Multi-factor measurements
  - E.g. patient data – age, blood pressure, pulse

Age	Blood pressure	Weight
24	120	65
48	140	100
27	130	70
32	90	55

=



Most data is multidimensional

- What about image data?

- What is the dimension of a 1 pixel gray image?

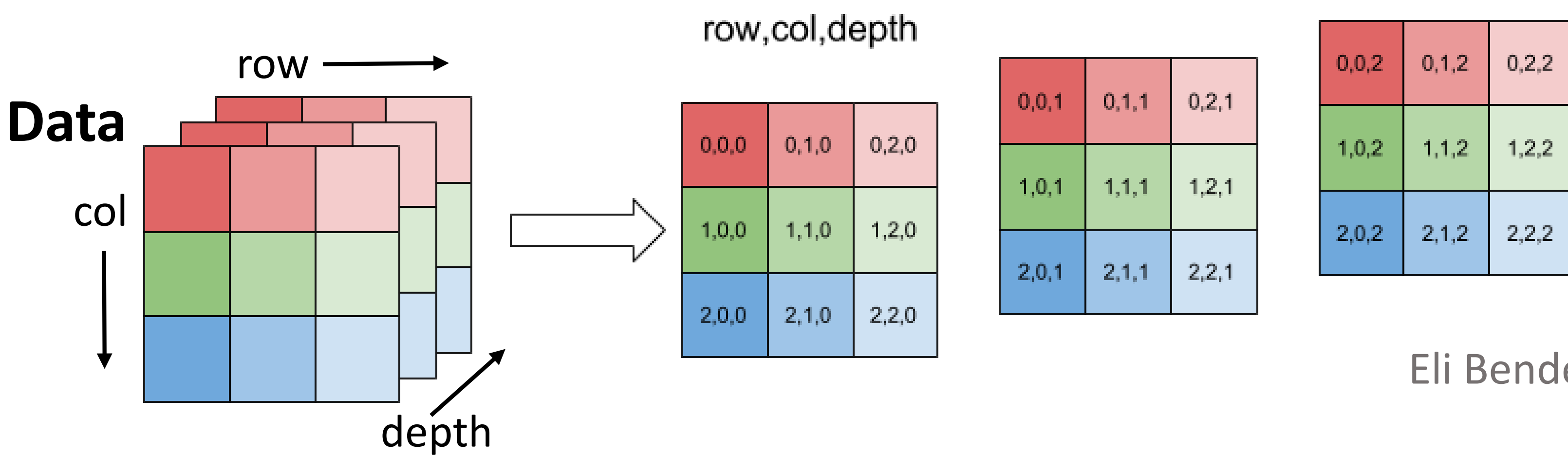
- Dimension $\cong$ the minimum number of coordinates needed to specify a point within a space

Most data is multidimensional

- Dimension ≅ the minimum number of coordinates needed to specify a point within a space
- In "image" space, each point is a different image
- What is the image-space dimension of a 100 x 100 grayscale image?

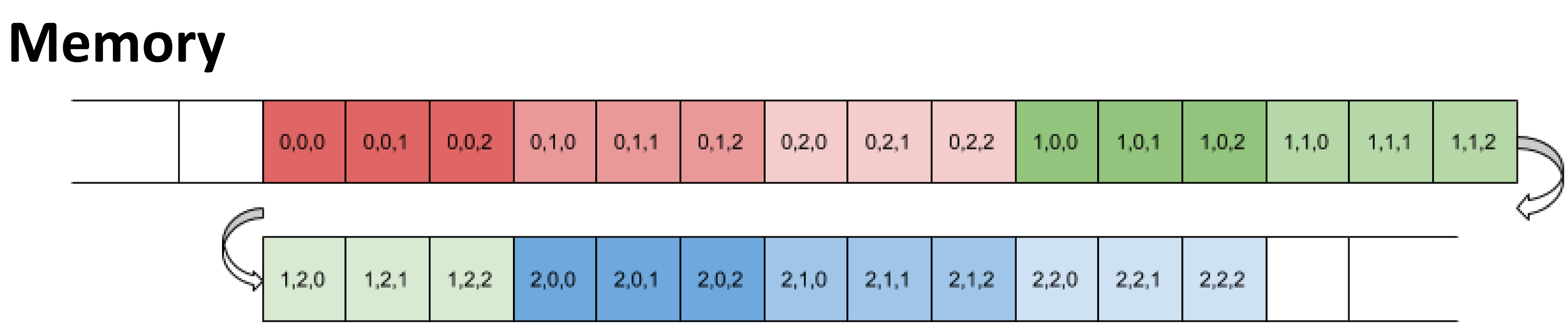
10,000 dimensions

Numpy n-dimensional arrays



Eli Bendersky

- Data is ‘wrapped’, in row major / C order



Eli Bendersky

xarray

- What about labeled arrays?
- Pandas in multiple dimensions?
  - Works, but dimension labeling and access gets awkward
- **xarray** – multi-dimensional pandas
  - Dimension names (dim=‘time’ instead of axis=3)
  - **DataArray** – labeled n-dim array (≅ pandas.Series)
  - **Dataset** – aligned DataArrays (≅ pandas.DataFrame)
  - Compatibility with Pandas, netCDF, dask
- *<notebook intro>*

Challenges of dimensionality

- Unintuitive mathematical features
- Visualization is difficult
- Computational costs/complexity
  - Worst case complexity: x^{n_dim}

Unintuitive math features

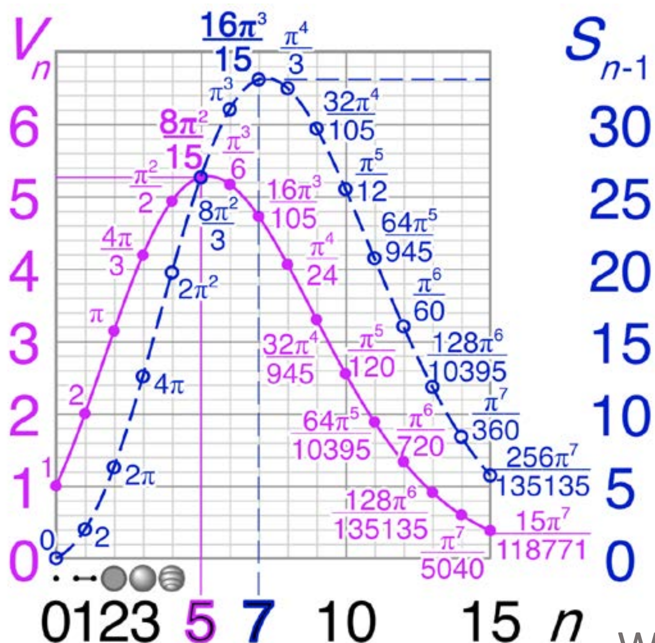
- Volume of hypercube = length^{ndim}
- Volume of hypersphere => it's complicated
  - For odd dimensions (1, 3, 5…):

$$Volume_{ndim}\,(radius) =$$

$$2 * \left(\frac{ndim-1}{2}\right)! \; * (4 * pi)^{(ndim-1)/2} \Big/ {ndim! * radius^{ndim}}$$

Unintuitive math features

- Hypersphere **volume decreases** with high dimension!

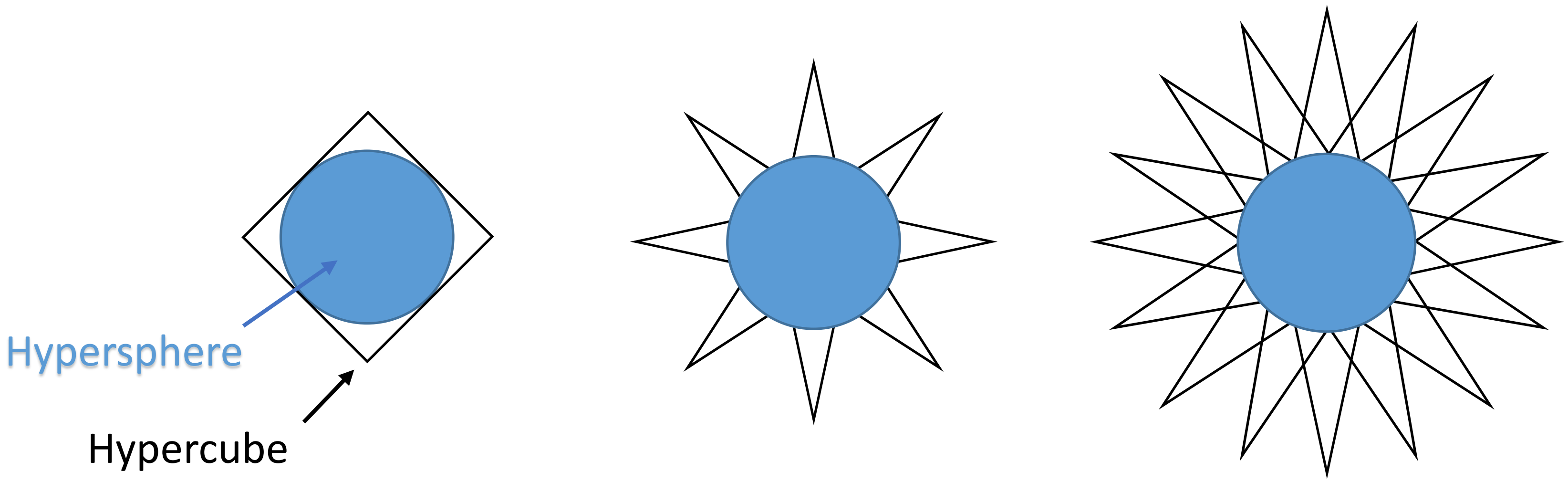


Wikipedia

$$2 * \left(\frac{ndim - 1}{2}\right)! \; * (4 * pi)^{(ndim\,-1)/2} \Big/ _{ndim!} * radius^{ndim}$$

Unintuitive math features

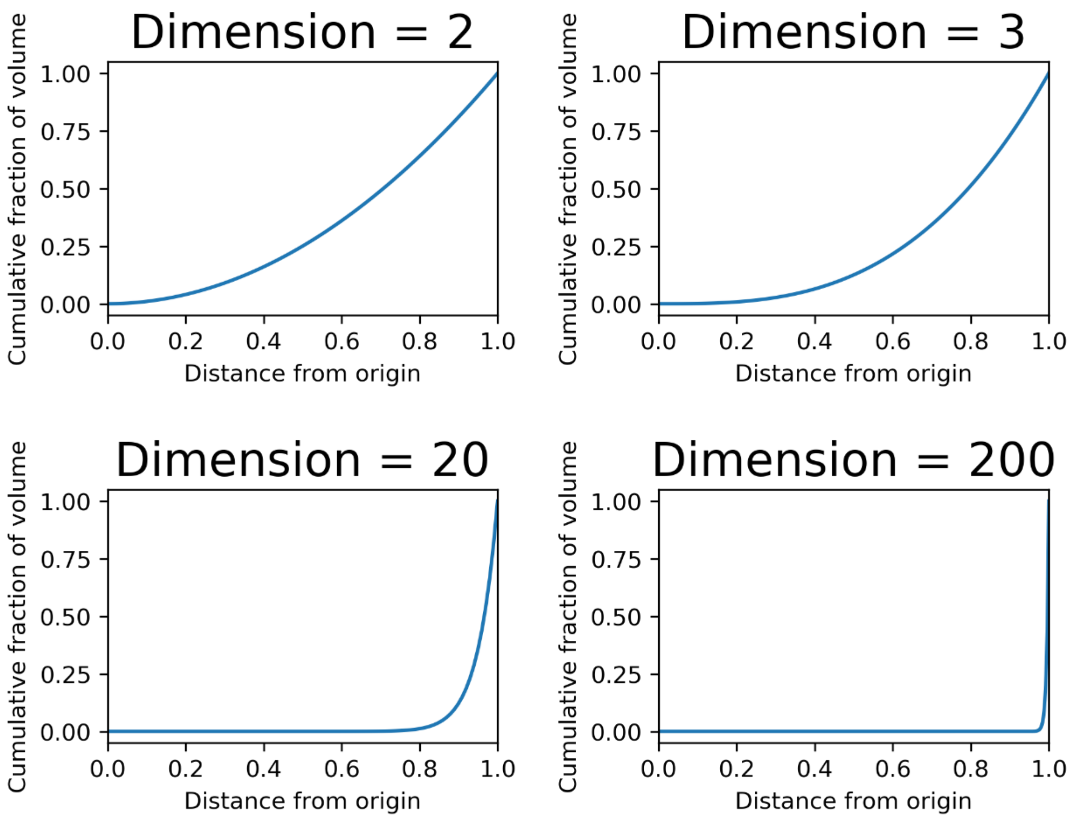
- It’s not that the radius changes
- Volume gets *weird*



- Volume becomes concentrated in the outer ‘skin’
  - Because of the $radius^{ndim}$ term

Unintuitive math features

- Volume becomes concentrated towards corners and outer ‘skin’

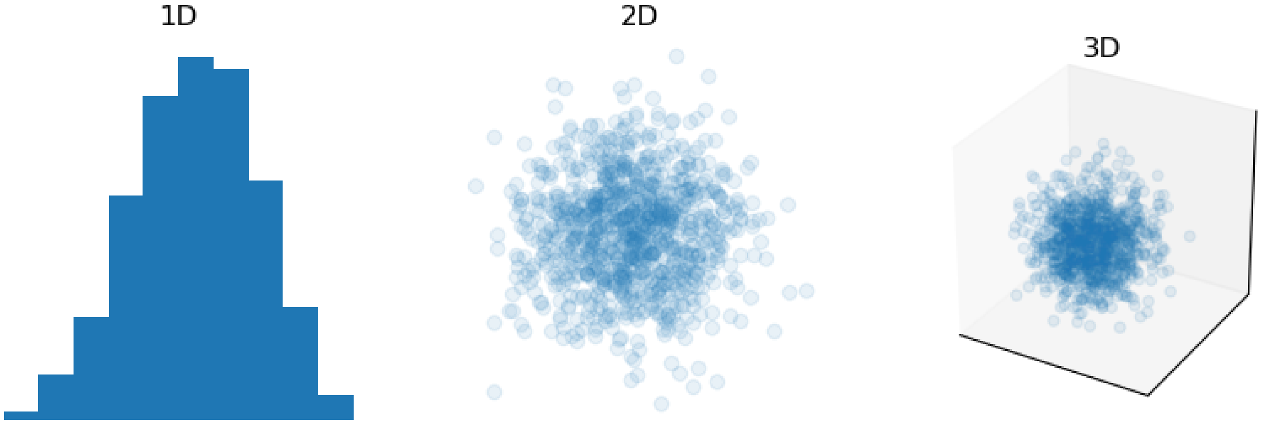


Unintuitive math features

- At higher dimensions:
  - Spheres have little volume
  - Volume becomes concentrated in the corners and skin
  - Small changes in radius/length change volume greatly
- Practical impact:
  - Few ‘nearby’ points (using Euclidean distance)

Unintuitive math features

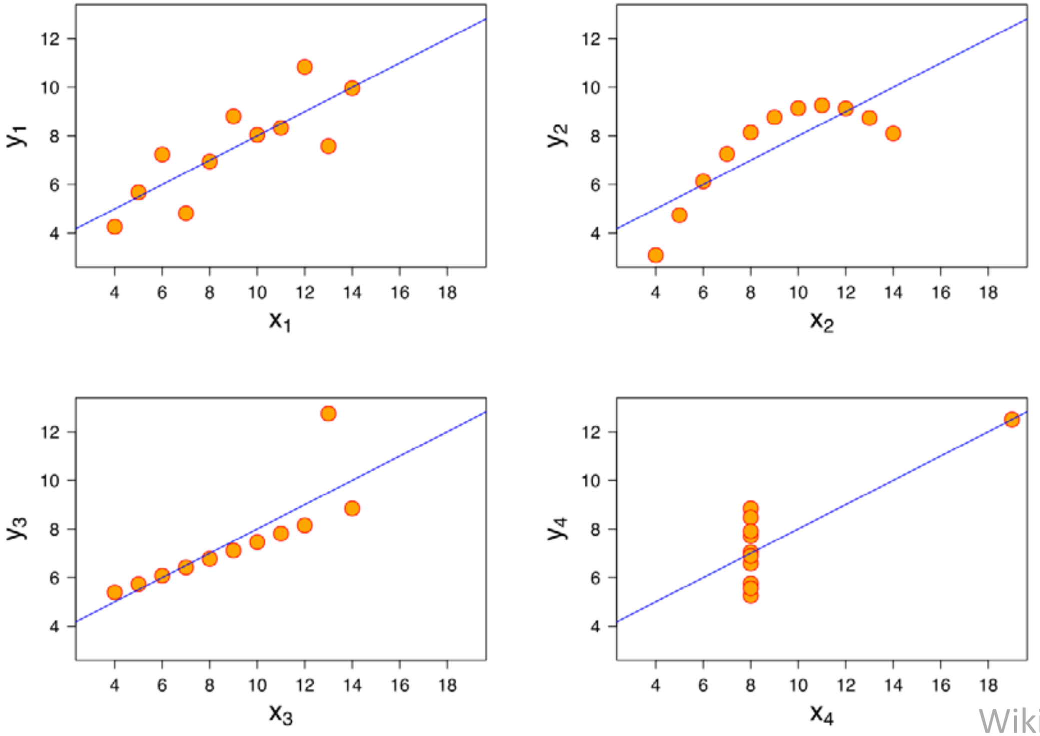
- Gaussian/normal distribution in high dimensions?



- Density is still always highest at the center
  - But there's not much volume in the center
- Probability mass becomes concentrated at the skin
  - Distributions become more 'bubble' like

Visualization issues

- Why is visualization important?



Wikipedia

All four have same mean and variance!

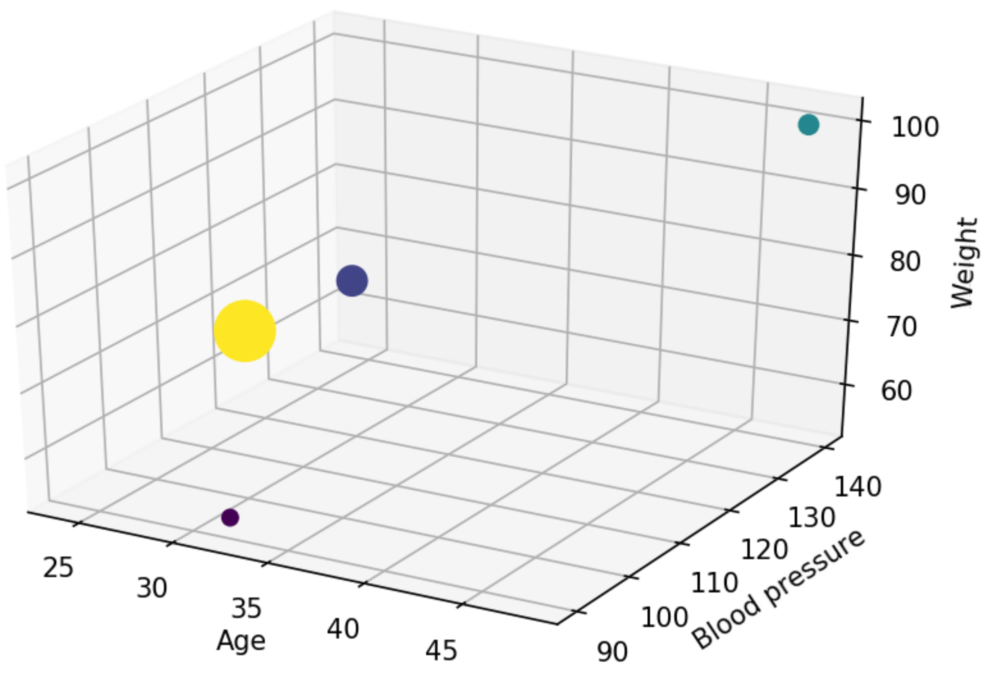
Visualization issues

- Why is visualization important?

Mean image (from 1000 samples)

Visualization strategies

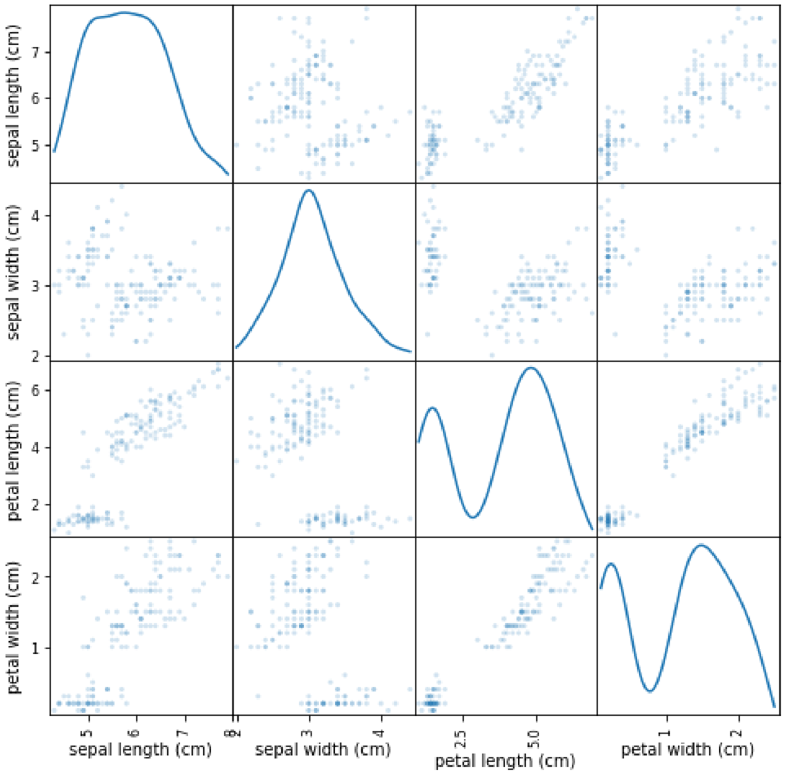
- ‘Overload’ with color, size, and shape



- Best for sparse data

Visualization strategies

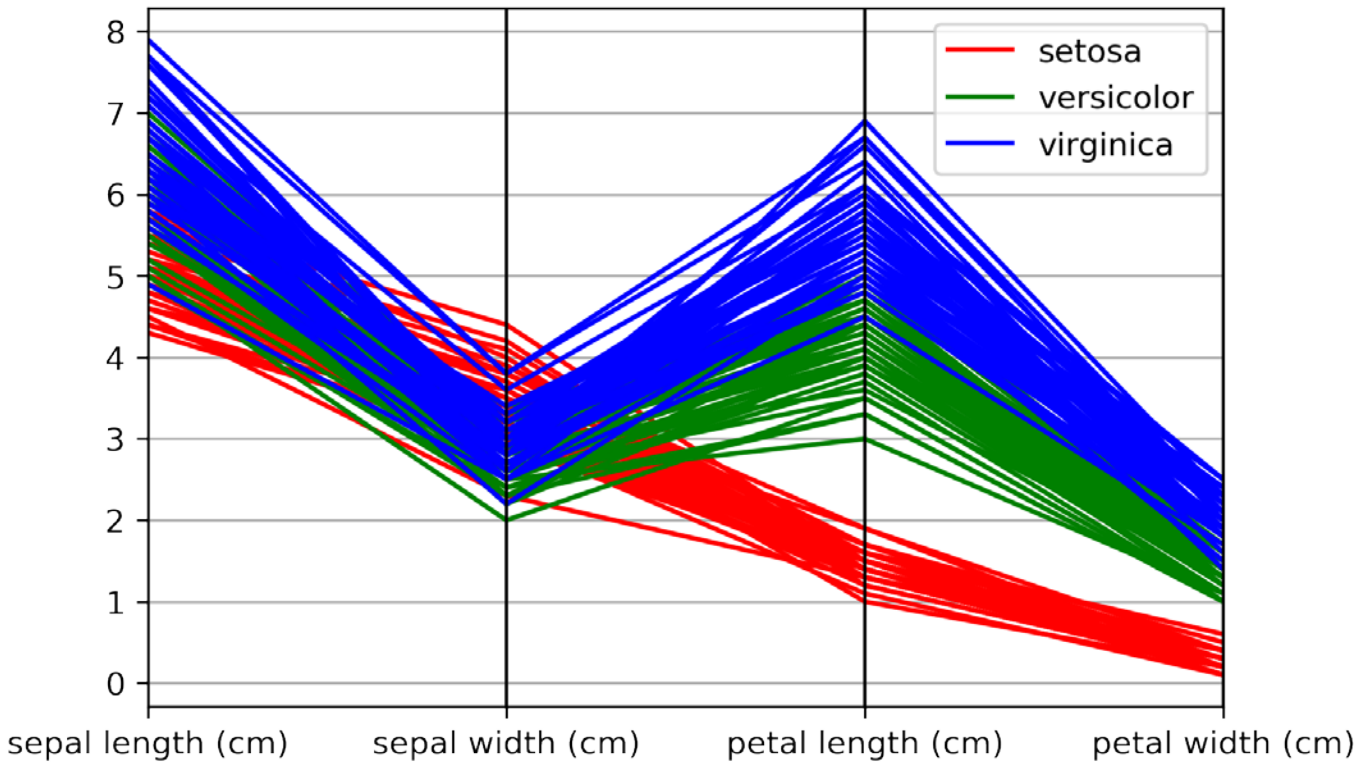
- Scatter plot matrix



from pandas.plotting import scatter_matrix

Visualization strategies

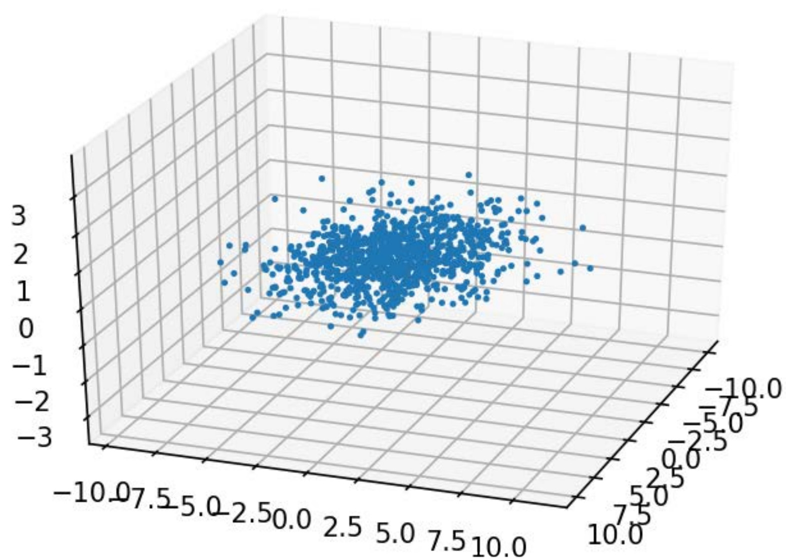
- Parallel coordinate plot



from pandas.plotting import parallel_coordinates

Dimensionality reduction

- Why?
  - Less dimensions can be nicer to work with
- Justification:
  - Often data isn't fully distributed in it's n-dimensional space
  - Equivalently – correlations in the data

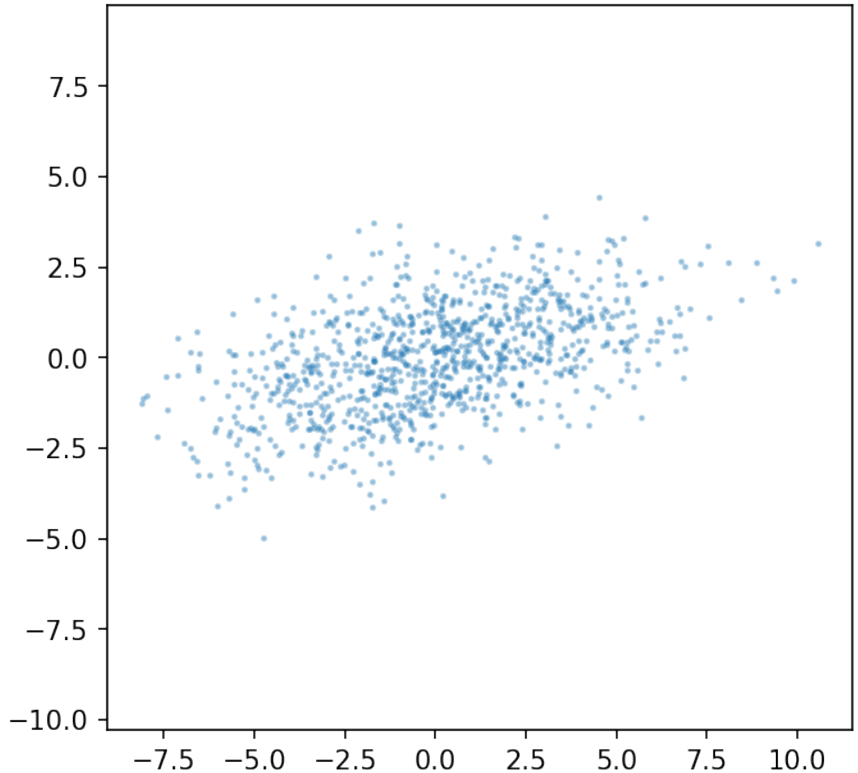


PCA – Principal Component Analysis

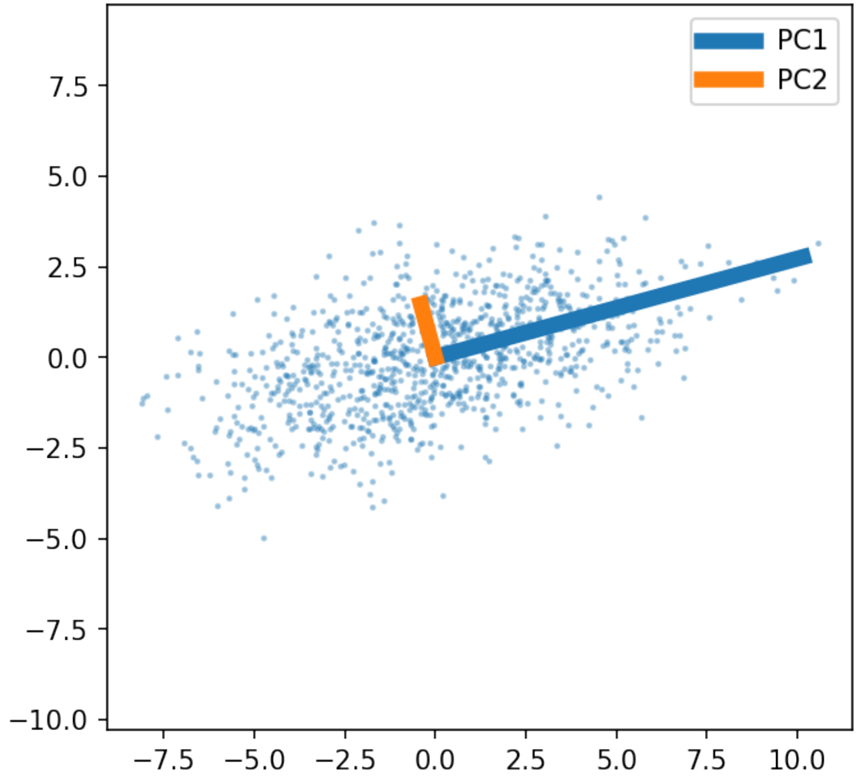
- Isn't just dimensionality reduction

- Can be thought of as:
  - Capturing covariance
  - Fitting an ellipsoid to the data
  - Rotation + scaling
    - Read up on SVD for details
  - Eigenvectors of the covariance matrix

PCA – Principal Component Analysis



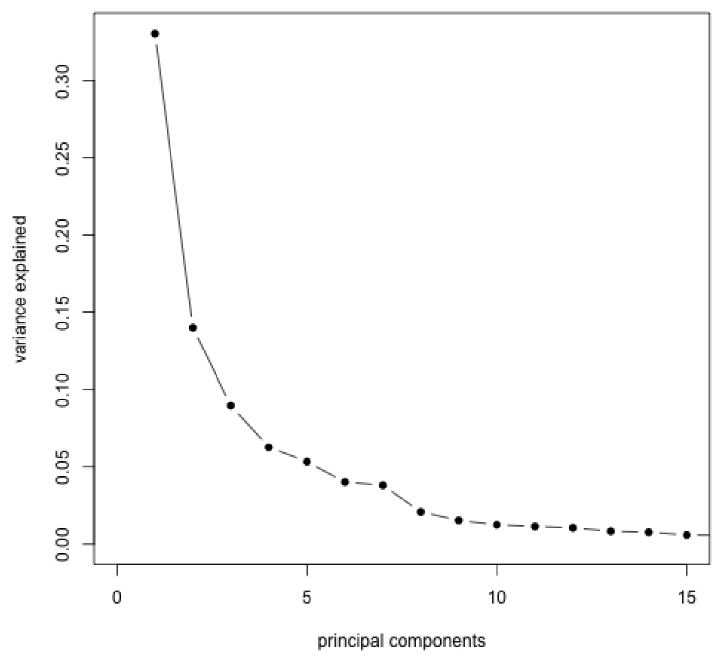
PCA – Principal Component Analysis

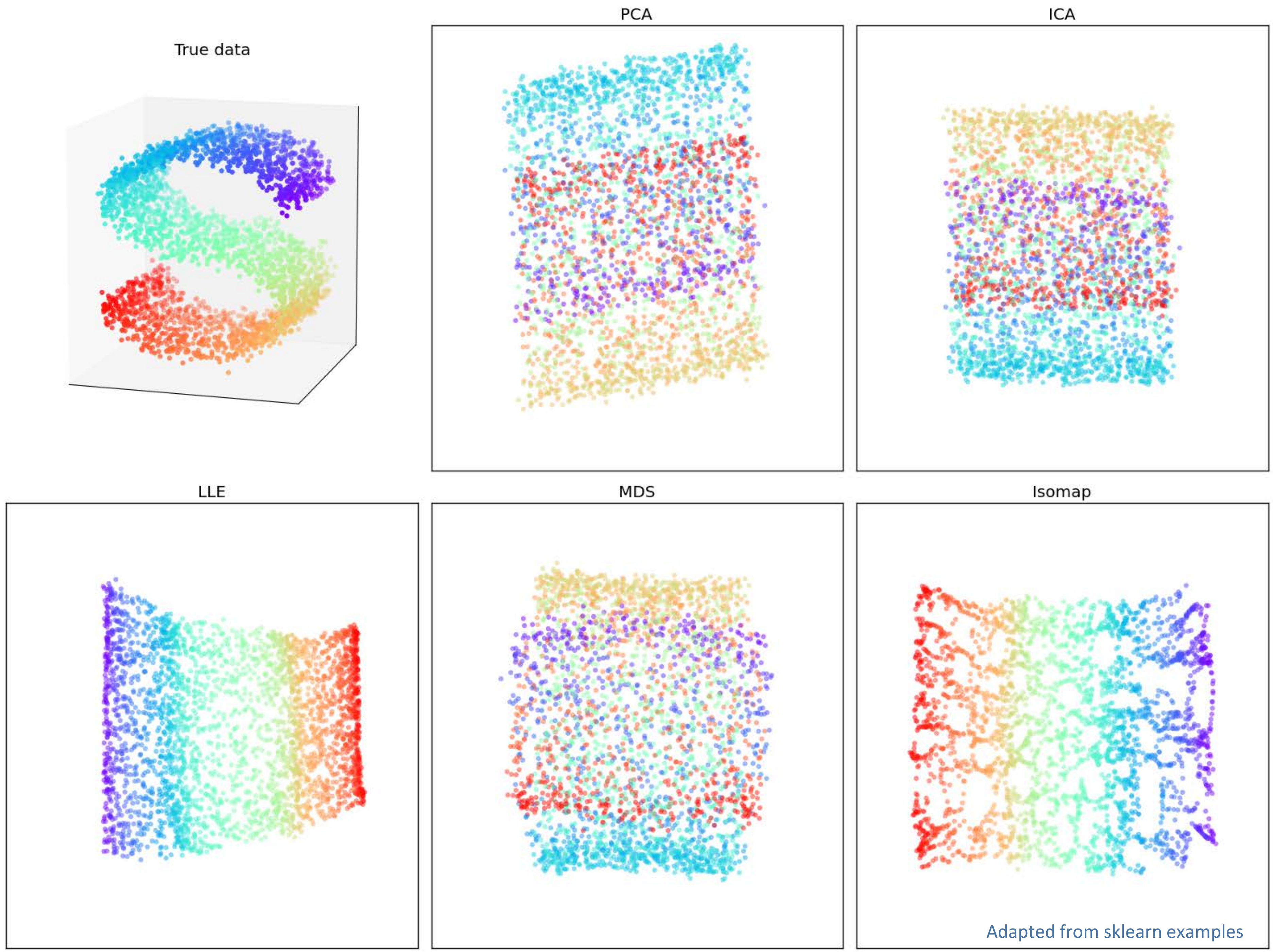


- Output: Principal Component (PC) vectors
- PCs always orthogonal

PCA

- Dimensionality reduction – truncate # PCs
- Explained variance
    - fraction of total variance explained per component

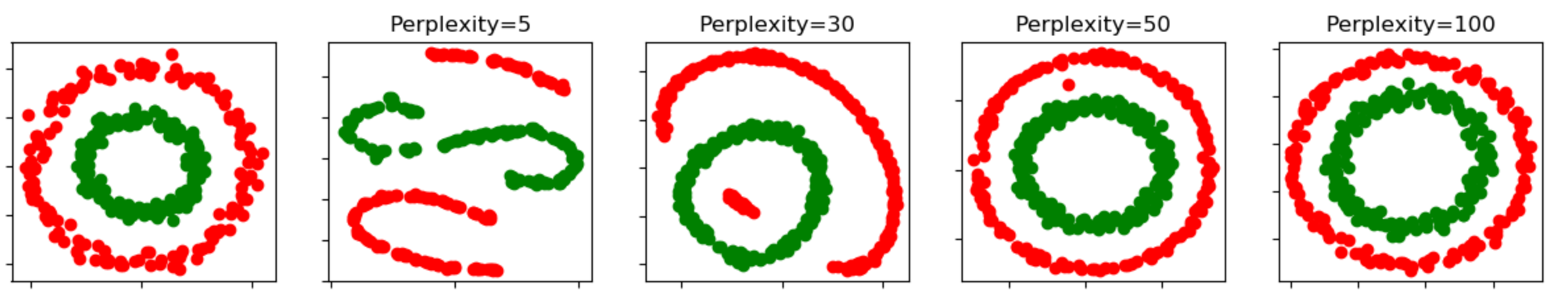




True data
PCA
ICA
LLE
MDS
Isomap
Adapted from sklearn examples

Visualization: t-SNE

- t-distributed Stochastic Neighbor Embedding
- Tries to preserve local structure
- Perplexity” parameter balances local and global



Sklearn

- Cluster sizes/shapes not meaningful
- Different random seeds can change output!
- Not ideal for use beyond visualization

Thanks!

- Thanks to JetBrains for hosting/sponsoring
- Deeper questions/discussion – come find me later!
- Our lab is always looking for programmers and those interested in computation + neuroscience:
  - joe@neuro.mpg.de

Misc. further references

[High dimensional spaces]

[t-SNE]

[PCA explained variance]